Conservation Biology

Letter

Using machine learning to advance synthesis and use of conservation and environmental evidence

Rapid growth in environmental research (Li & Zhao 2015) presents a potential wealth of information for use in conservation decision making. Evidence synthesis methods (e.g., systematic maps, reviews, meta-analyses) (Pullin & Knight 2009) are critical for garnering actionable insight from published research, yet they require levels of time and funding that are prohibitive for meeting short policy windows (Elliott et al. 2014) and balancing trade-offs between conservation planning and implementation.

In response, interest in machine learning to make syntheses faster and more efficient is growing (O'Mara-Eves et al. 2015). Machine learning (ML) is based on the idea that computers can be programmed to automatically perform a set of tasks by learning from a set of rules and training data (Alpaydin 2014). For example, ML could be used to synthesize information by semiautomatically finding articles relevant to users and even to summarize information—potentially reducing time and bias and improving overall cost-effectiveness. Machine learning has been widely applied in public health and syntheses of medical information and is beginning to be trialed in conservation and environmental topics (Westgate et al. 2015; Roll et al. 2017). Bearing the challenges in mind, we endeavored to design a platform, powered by machine learning, to improve on the manual synthesis process. We partnered with DataKind, a data-science nonprofit organization, to create an open-access platform, Colandr, to address 2 laborious stages of information synthesis: finding relevant articles and extracting desired information. Colandr has 2 learning systems, the first iteratively sorts articles by relevance as specified by users and the second aids in categorizing article topics (Fig. 1). To illustrate Colandr's functionality, we used data from a systematic map on linkages between conservation and human well-being (McKinnon et al. 2016).

Reviewers typically sort through thousands of search results to find relevant articles, an inefficient process that often takes several months. For example, our search on conservation and human well-being recovered 35,000 hits, of which only 1,000 were relevant. System 1 in Colandr aims to speed up this process by dynamically sorting the wheat from the chaff based on user input. As users indicate whether citations are relevant or irrelevant, system 1 calculates the expected relevance of the remaining search results and dynamically pushes more relevant citations to the top. Colandr does this by searching for patterns in the words around search terms (e.g., it identifies the words and the order of those words around *protected area*) and learning which of these combinations are more relevant to the user, a method called word2vec (Mikolov et al. 2013). System 1 achieved a 5-fold reduction in effort with our systematic map data set. Manually, reviewers had to read 1,436 citations before they recovered 100 relevant articles. Using Colandr, reviewers recovered 100 relevant citations after reading only 250 citations.

After finding relevant studies, reviewers embark on pulling out desired information (e.g., bibliographic, topical, results) from each article. For example, we categorized articles according to topic area (e.g., types of conservation and human well-being). Typically, reviewers read entire articles to categorize them, a very time-consuming process. System 2 is designed to deduce these categories faster by pulling sentences from articles that it identifies as relevant to each category with the global vectors for word representations (GLoVE) model (Pennington et al. 2014). For example, Colandr will display sentences related to the law and policy category for users to read and help them decide whether to accept or reject that category. As users continue to categorize articles, model confidence improves. Although this approach does not necessarily improve speed, it can improve accuracy by catching missing or mislabeled categories.

Colandr semiautomates the synthesis process, but it continues to retain significant user oversight to ensure transparency. This is critical because conservation and environmental terms often have alternative meanings and many synonyms. For example, there are many different types of protected areas (e.g., key biodiversity areas, reserves, no-take zones, biospheres, national parks), whereas there are only 2 ways to refer to influenza, *influenza* or *flu*. This heterogeneity makes it harder (but not impossible) to make automatic predictions with high

***Article impact statement**: Machine learning optimizes processes of systematic evidence synthesis and improves its utility for evidence-based conservation.*

Conservation Biology, Volume 32, No. 4, 762–764

DOI: 10.1111/cobi.13117

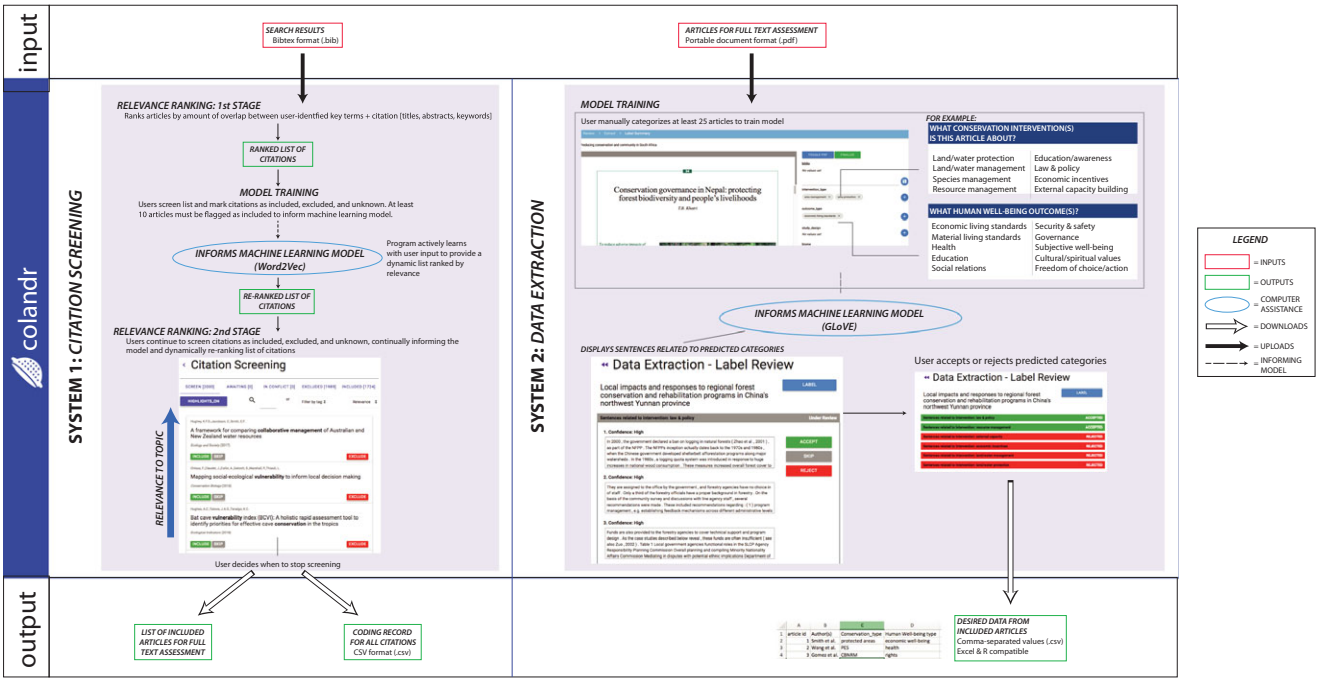

Figure 1. Colandr is composed of 2 systems in which user input iteratively trains the machine-learning models. In system 1, articles are ranked by relevance. In system 2, topic labels are predicted for each article. Arrows illustrate the flow of inputs and outputs during a systematic evidence synthesis project and where model training occurs.

levels of confidence. Thus, because this process is less precise, Colandr aims to retain user oversight to ensure that relevant articles are not missed.

Preliminary tests of Colandr demonstrate significant improvements over a manual process. Although such assessments have not been exhaustive, they demonstrate Colandr's potential to help advance evidence-based decision making in conservation by removing resource barriers to conducting comprehensive evidence syntheses.

Acknowledgments

We thank the Science for Nature and People Partnership (SNAPP), a partnership of The Nature Conservancy, the Wildlife Conservation Society, and the National Center for Ecological Analysis and Synthesis (NCEAS) at the University of California, Santa Barbara, for providing funding for the Evidence-Based Conservation working group. Development and deployment of Colandr was conducted in partnership with DataKind and Conservation International. We especially thank S. Sagalovsky for front-end development of Colandr and E. Fegraus at Conservation International. Finally, we thank A. Pullin, M, Balisi, A. Fritts-Penniman, and S. Bittick for providing comments on earlier drafts of this manuscript. R.G. is partially supported by the National Institute for Health Research (NIHR) Collaboration for Leadership in Applied Health Research and Care (CLAHRC) for the South West Peninsula (PenCLAHRC).

S.H. Cheng,[1] C. Augustin,[2] A. Bethel,[3] D. Gill,[4,5] S. Anzaroot,[2] J. Brun,[1] B. DeWilde,[2] R.C. Minnich,[2] R. Garside,[6] Y.J. Masuda,[7] D.C. Miller,[8] D. Wilkie,[9] S. Wongbusarakum,[10] and M.C. McKinnon[11]

[1]National Center for Ecological Analysis and Synthesis, University of California, Santa Barbara. 735 State Street, Suite 300, Santa Barbara, CA 93101, U.S.A., email samantha.cheng@asu.edu
[2]DataKind, 156 5th Avenue, Suit 502, New York, NY 10010, U.S.A.
[3]University of Exeter Medical School, Heavitree Road, Exeter, EX1 2LU, U.K.
[4]Conservation International, 2011 Crystal Drive, Suite 500, Arlington, VA 22202, U.S.A.
[5]Environmental Science and Policy, George Mason University, Fairfax, VA 22030, U.S.A.
[6]European Centre for Environment and Human Health, University of Exeter, Truro, Cornwall, TR1 3HD, U.K.
[7]The Nature Conservancy, 4245 Fairfax Drive, Arlington, VA 22203, U.S.A.
[8]Department of Natural Resources and Environmental Sciences, University of Illinois, S-406 Turner Hall, 1102 S. Goodwin Avenue, Urbana, IL 61801, U.S.A.
[9]Wildlife Conservation Society, 2300 Southern Boulevard, Bronx, NY 10460, U.S.A.
[10]Social Science Research Institute, University of Hawaii, 2424 Maile Way #704, Honolulu, HI 96822, U.S.A.
[11]Vulcan, Inc., 505 Fifth Avenue S, Seattle, WA 98104, U.S.A.

Literature Cited

Alpaydin E. 2014. Introduction to machine learning. MIT Press, Cambridge.

Elliott JH, Turner T, Clavisi O, Thomas J, Higgins JPT, Mavergames C, Gruen RL. 2014. Living systematic reviews: an emerging opportunity to narrow the evidence-practice gap. PLoS Medicine **11**(2):e1001603. https://doi.org/10.1371/journal.pmed.1001603.

Li W, Zhao Y. 2015. Bibliometric analysis of global environmental assessment research in a 20-year period. Environmental Impact Assessment Review **50:**158–166.

McKinnon MC, et al. 2016. What are the effects of nature conservation on human well-being? A systematic map of empirical evidence from developing countries. Environmental Evidence **5:**8.

Mikolov T, Sutskever I, Chen K, Corrado G, Dean J, Chen K, Dean J, Mikolov T, Chen K. 2013. Distributed representations of words and phrases and their compositionality. NIPS'13 Proceedings of the 26th International Conference on Neural Information Processing Systems **2:**3111–3119.

O'Mara-Eves A, Thomas J, McNaught J, Miwa M, Ananiadou S. 2015. Using text mining for study identification in systematic reviews: a systematic review of current approaches. Systematic Reviews **4:**5.

Pennington J, Socher R, Manning CD. 2014. GloVe: global vectors for word representation. Pages 1532–1543. Proceedings of the 2014 Conference on empirical methods in natural language processing (EMNLP). Association for Computational Linguistics, Doha, Qatar.

Pullin AS, Knight TM. 2009. Doing more good than harm - Building an evidence-base for conservation and environmental management. Biological Conservation **142:**931–934.

Roll U, Correia RA, Berger-Tal O. 2017. Using machine learning to disentangle homonyms in large text corpora. Conservation Biology **32:**716–724.

Westgate MJ, Barton PS, Pierson JC, Lindenmayer DB. 2015. Text analysis tools for identification of emerging topics and research gaps in conservation science. Conservation Biology **29:**1606–1614.